International Journal of Civil Engineering and Technology (IJCIET)
Volume 8, Issue 3, March 2017, pp. 468–478 Article ID: IJCIET_08_03_047
Available online at http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=3
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

© **IAEME** Publication **Scopus** Indexed

IDS FEATURE REDUCTION USING TWO ALGORITHMS

Safana H. Abbas
Assistant Professor
Computer Science, College of Education,
AL-Mustansiriya University, Baghdad, Iraq

ABSTRACT

In recent years, intrusion is defined as detection of any security threats .The security of the information has become a very dangerous problem in the security of the data and network. Highly secret data of different arrangements are present via the network so in order to protect that data from unauthorized users, it is required a very strong security structure. An IDS (Intrusion detection system) gathers and tests information from various areas within a network to determine the most likely security threats that from both outside and inside the system. IDS deals with huge data which include different redundant and irrelevant features that results in increasing time processing and decreasing detection rate. Therefore reduction of features plays an important role in IDS. In this paper two dimensionality reduction algorithms PCA and SVD were implemented on KDDCUP'99 dataset. Experimental results were obtained to get the best reduced feature set that recognized using SVM algorithm. Detection rate, error and accuracy are used to evaluate IDS Performance.

Key words: Intrusion Detection, Kddcup99 Dataset, Feature Reduction, Classification.

Cite this Article: Safana H. Abbas, IDS Feature Reduction Using Two Algorithms. *International Journal of Civil Engineering and Technology*, 8(3), 2017, pp. 486–478.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=3

1. INTRODUCTION

Many intrusion detection systems are tested and proposed in the last few years to overcome the internet that is vulnerable [1]. Depending on the results of" American Computer Emergency Response Team /Coordination Center (CERT)" [2], in recent years networking showed great index increasing and have became the world war's new weapon [3]. Furthermore the report said that "Chinese Military Hacker" had made a plan depending on the attacking" American Aircraft Carrier Battle Group "view to be a weak fighting range via internet. These information leads to a quick need that determine and prevent internet attacks [4]. Therefore we can say that an IDS is very important for new computer systems. There are two general types to computer IDS which are anomaly detection and misuse detection. The misuse detection is when a known attack signature is recognized an alarm is generated, while Anomaly detection determines an event that different from the regular attitude of the

observed system and therefore new attacks can be recognized [5]. The information that used is coming from "MIT's Lincoln Lab". It was organized for KDD (Knowledge Discovery and Data mining) race by DARPA and it is used a basic evaluations for ID program [6]. Experimental tests concludes that the algorithms of feature reduction can reduce the dataset size. The space and time difficulties of the most used classifiers are" exponential function of their input vector size" [7]. Further, the need for the patterns and numbers for training the classifiers increased exponentially with the volume of the features space. This restriction is called the "curse of dimensionality". The feature space having decreased features that correlates to classifications and minimizes the costs of the pre-processing and decreasing the effects of the classification peaking event [8].

Today the computer and internet have become an important issues in our life. The internet openness and scalability have made it adaptable stage for an on-line services new generation, such as military, E-commerce, public web services, social network, online shopping, stock prices, etc. The publicity of these services caused in a large financial volume that deals with secret information being accessed through the internet. Internet has high various security subjects because of the huge use of network, the value and importance of this information and the correlated on-line services which have made the internet a council for a wide different types of attacks [9, 10].

2. NETWORK SECURITY

It consists of the policies and conditions that are adopted by an administrator of the network to stop and detect misuse, unauthorized access, modification of a computer network and resources of the network is accessible [11].

2.1. Intrusion and Intruder

Intrusion: It is the violation into a computer system or network and badly using them to perform the virulent activities. When an information system user takes an activity where the user is not legally allowed to use is called *Intrusion* [11].

Intruder: Is the individual who attacks the computer system or network and badly using the computer system or network is called as an *Intruder*. two kinds of Intruders are existed called internal intruder and external intruder [11]. *Internal Intruder* is a person who override his bounded authority to made an action. His action may or may not be hurts the system or the services provided by the system but it requests to earn extra capability to made an action without allowable authorization. [13, 14].

External Intrusion comes from outside of the system and harming computer system or network. *External intruders* do not have any legally accessing to the system they attack. An example of *external intruders* are hackers [12].

2.2 Intrusion Detection System (IDS): It is the mechanism of observing and analyzing the events happened in a computer system to detect signals of security troubles, that helps in determining a set of strange actions that arranges the "integrity, confidentiality and availability of information resources". ID is a complex issue because of the main thought of detection speed, detection accuracy, the dynamic circumference of the networks and the processing power for processing huge data from segmented network systems [15].

2.3 Detection Methodologies

ID methodologies are categorized in two main classes:

2.3.1 Signature-based Detection (SD): An SD is a string that relates to a known attack. Signature-based Detection is the process to match a pattern against recognized intrusions. Because of used information gathered by certain attacks and system intruders, Signature-based Detection is also known as "Knowledge-based Detection or Misuse Detection" [16, 17].

2.3.2 Anomaly-based Detection(AD) : An AD is a variation from a network connections, known behavior that detect the expected behaviors deviated from monitored regular activities, host or users over an interval of time [18, 19].

3. PROPOSED SYSTEM

The proposed system is consisting mainly of two major tasks which are:

1. Feature Reduction.
2. Attack Detection.

The proposed intrusion detection system is illustrated in figure (1) which consisting of the following stages:

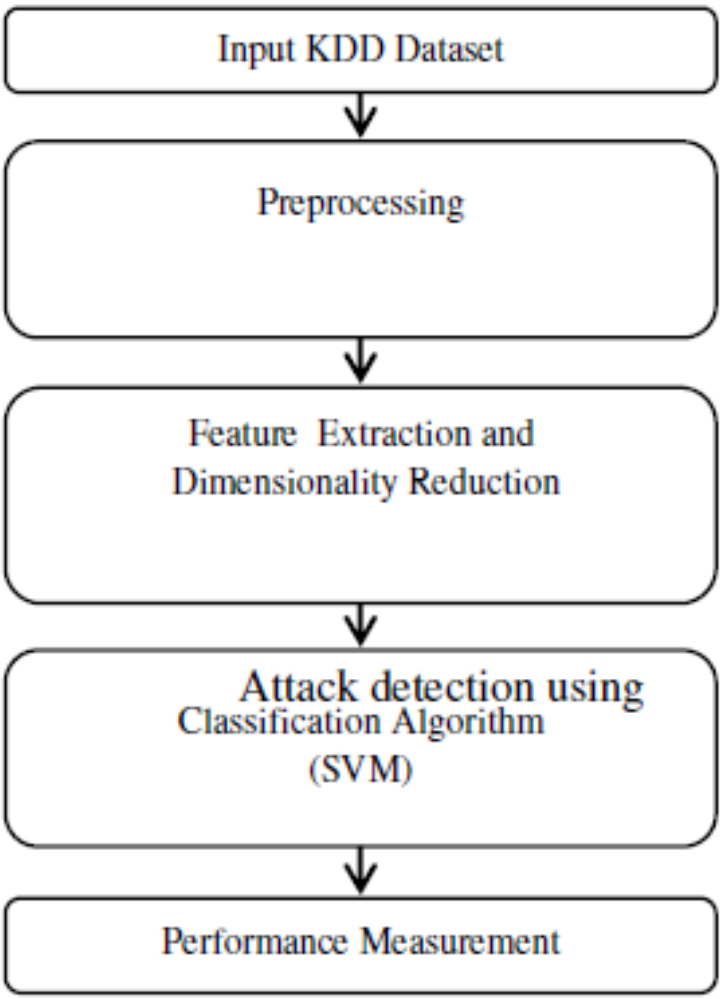

Figure. 1 Proposed IDS.

A. Preprocessing stage

1. Labeling: The dataset should be labeled by using 10% of the corrected dataset of the whole feature space. Every record in the dataset contains 42 features (e.g., protocol type, service, and Flag) and is labeled as either normal or an attack with one specific attack type as shown in Figure (2), which is a sample from the dataset before normalization, first row as an example. It can be noticed that the feature (42) has the normal type of attack.

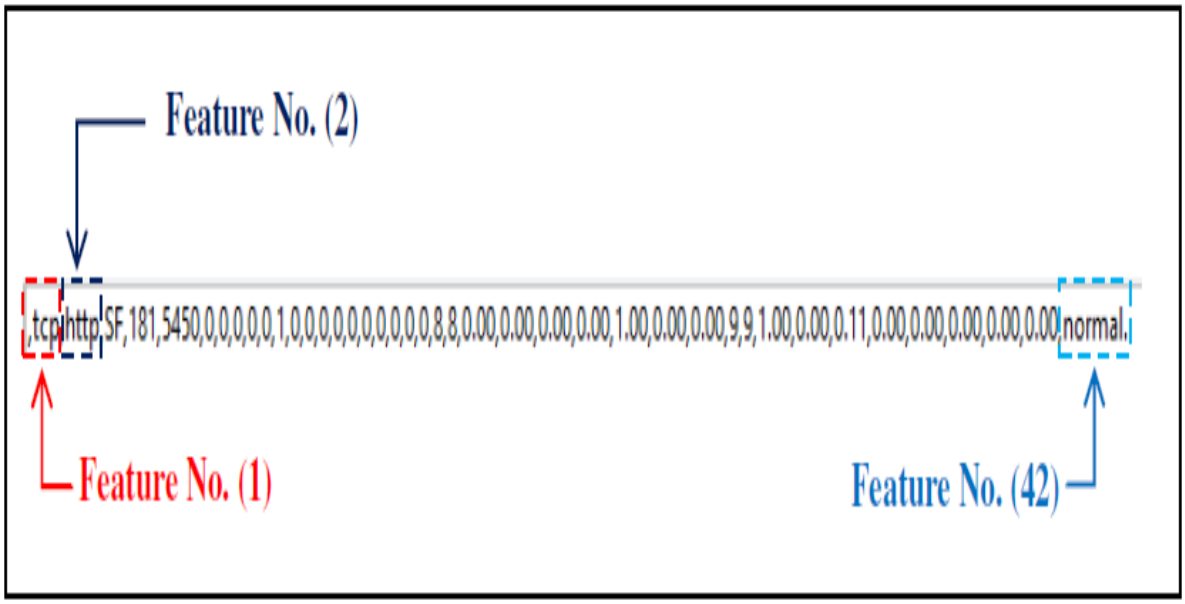

Figure 2 First row (data sample) of 10% correction KDD cup dataset

2. Normalization: It is used where the attribute data are scaled so as to fall within a small specified range such as (-1 to 1) or (0 to 1). Normalizing the input values for each attribute measured in the training samples will help in speeding up the classification stage.

B. Dimensionality reduction stage

Two different algorithms principle component analysis (PCA) and singular value decomposition (SVD) are used to reduce the 42 features as much as possible and applying these reduced features to the recognition algorithm later:

1. Principal component analysis (PCA)

PCA is a helpful statistical method. Its main goal is to decrease the data dimension while keeping the different present in the original dataset.

Algorithm (1) PCA

Input: Proposed Trained dataset.
Output: PCA set of most frequent and related features.
Steps:
1. acquire training KDD transactions.
2. Represent each transaction (Ii) as a vector (xi).
3. The average transaction is computed
4. The mean transaction is subtracted
5. Estimate the covariance matrix =AAT
6. estimate eigenvectors(ui)of AAT:
a. Consider AAT as a matrix.
b. Estimate the eigenvectors(vi) of (AAT) such that:
ATAvi → iVi → AATAVi = iAvi → Cui = iui where i =Avi
c. Estimate the best eigenvectors of AAT: i= Avi
7. Save only K eigenvectors.

2. Singular Value Decomposition (SVD)

SVD permits an accurate representation of any matrix, and remove the less significant segments of that representation to deduce an approximate representation with any required dimensions number. removing the least important items gives a smaller representation that nearly approximates the original matrix

Algorithm 2 : SVD
Input: Generate Data matrix X

Output: New Dimensions C

1. **Repeat**
2. Applying SVD to the matrix X as $X = USV^T$
   $X \rightarrow$ is an $m \times n$ matrix

   $-m \rightarrow$ no. of sessions (vectors)

   $-n \rightarrow$ is no. of attributes)

   $U \leftarrow XX^T$ matrix of the eigenvectors

   S is matrix which is diagonal

   $V \leftarrow$ is matrix the eigenvectors.
3. Construct the covariance matrix from this decomposition by
   $XX^T\ XX^T \leftarrow (USV^T)(USV^T)^T = (USV^T)(VSU^T)$
4. $V \rightarrow$ an orthogonal matrix $(V^TV = I), XX^T = US2U^T$
5. square roots of the eigenvalues of XX^T are the singular values of X
6. **until**Represent every transaction Ii over the time interval t as a vector $x(t)_i$

Return U^TX

C. Attack detection stage

Support vector machine (SVM) is the learning machine algorithm that can perform binary classification and regression estimation tasks. It is becoming increasingly popular as a new paradigm of classification and learning because of two important factors. First, unlike the other classification techniques, SVM minimizes the expected error rather than minimizing the classification error. Second, SVM employs the duality theory of mathematical programming to get a dual problem that admits efficient computational methods. Support Vector Machines (SVM), which is introduced by the linearly separable two class problems. The optimal separating hyper plane for such problems is considered. Deals with techniques to handle linearly inseparable two class problems. Discusses non-linear Support Vector Machines. Finally, states the universal approximation property of Support Vector Machines.

Algorithm (3): SVM

Input: Train D training dataset, Test D testing dataset that has not been recognized

Output: Test D testing dataset that has been recognized

Steps:

1. All points in training dataset are initialized as (Xi, Yj) where X is a data vector and Y is classes vector.
2. Vector of weight (W) is initialized.
3. All points (x, y) are distributed and the hyper plane separator is extracted.
4. If the hyper plane obtain optimal division then use the hyper plane to classify Test D and go End else do the following steps
5. Maximize the hyper plan
6. Initialize Large multiplier *αi vector α*
7. Use classification function
8. Determine the support vectors (xi) with non-zero *αi*
9. Use the hyper plan resulted after determining support vectors as the classifier model

   End

4. CRITERIA FOR EVALUATION

To evaluate the performance of the proposed model the Accuracy, detection rat, false alarm and confusion matrix are estimated, by calculating True Positive, True Negative, False Negative and False Positive, as illustrated below:

- Accuracy = TP + TN / TP + TN + FP + FN [1]
- Detection rate= TP/ TP + FP [2]
- False alarm= FP/ FP + TN [3]
- A confusion matrix that determines the number of samples predicted incorrectly or correctly by a classification model:

5. RESULTS

SVM) is a classification algorithm that is used for the proposed intrusion detection system, with feature dimensionality reduction algorithms (PCA and SVD) depending on the KDD'99 Cup datasets.

In the training part, the KDD 99 train dataset have (sample records), R2L and U2R attack classes have few patterns in their class, and also DOS and probe class and all other from remaining classes. Training is performed on full featured dataset as well as feature reduced dataset.

5.1. Principal Component Analysis (PCA)

The first algorithm that we used for the dimensionality reduction is the Principal Component Analysis (PCA). We used the PCA depending on different number of selected k (the reduced number of feature). By trial and error the values 7, 11 and 21 features from the original (42) feature were tested. Table (1) shows the evaluation results after applying the SVM classification algorithm on KDD Cup 99 testing datasets and PCA using (k=21).

Table (1) Evaluation of the proposed IDS using SVM and PCA with K=21 on testing dataset

Confusion matrix				Detection Rate	False Alarm	Accuracy
TP	FP	TN	FN			
98.1681	0.8319	99.3548	0.6452	99.1681	o.3833	99.2063

5.2. Singular-Value Decomposition (SVD)

The SVD is used depending on different number of selected features which is called k-feature section (the reduced number of features), usually after mean centering (normalizing) the data matrix for each feature. That means that there is no majority to give us any clue about the k selection so we test the SVD using three deferent k values (k=21, 11, and 7) from the original (42) features. Table (2) shows the evaluation results after applying the SVM classification algorithm on KDD Cup 99 testing dataset with k=21.

Table (2) Evaluation of the proposed IDS using SVM and SVD with K=21 on testing dataset

Confusion matrix				Detection rate	False alarm	Accuracy
TP	FP	TN	FN			
99.8841	0.1159	0	100	99.8841	100	99.9074

6. DISCUSSION

Table (3) shows the overall performances results of Support Vector Machine (SVM) on KDD Cup 99 depending on testing datasets by using two dimension reduction algorithms (PCA and SVD) when k=21 which gives the best accuracy than other K values.

Table (3) Accuracy of using SVM with SVD and PCA on testing dataset

Dimension reduction Algorithms	Features No.	Average accuracy
PCA	21	99.2063
SVD	21	99.9074

Figure (3) illustrate the performance results on the testing dataset using two dimensionality reduction algorithm that we used with support Vector Machine classification algorithm.

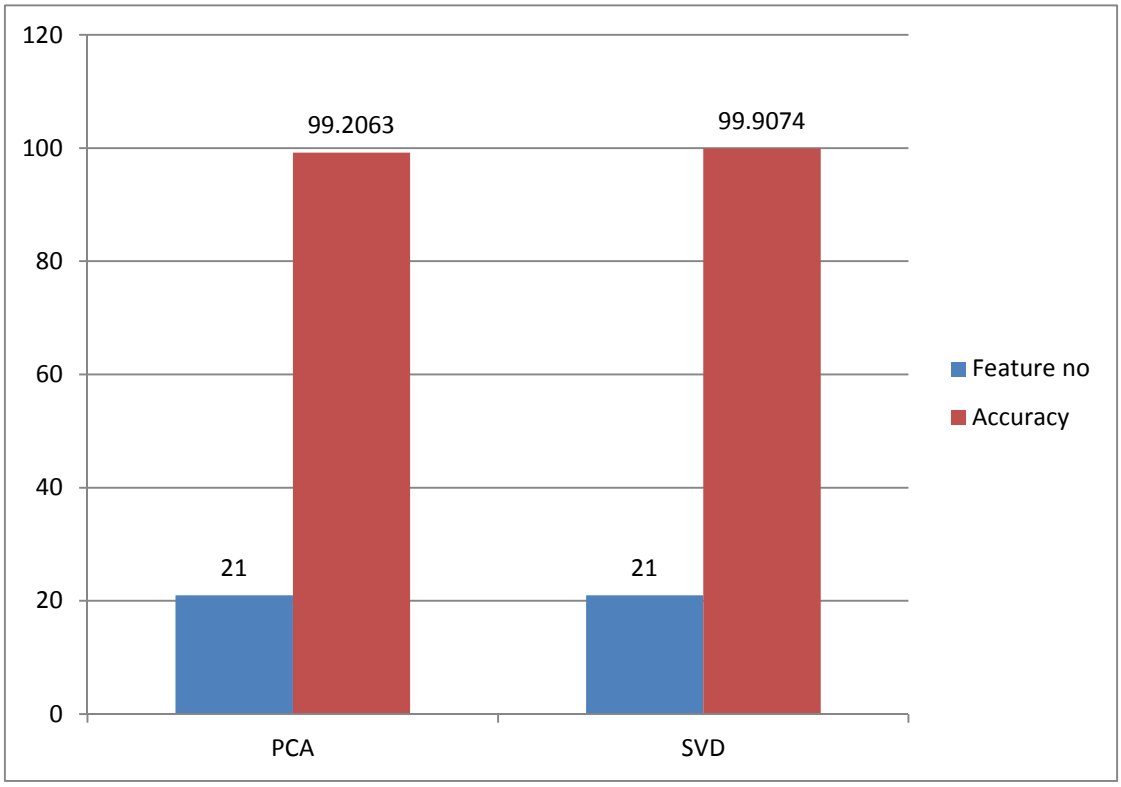

Figure (3) Accuracy of using SVM with PCA and SVD on training data

7. CONCLUSIONS

The aim of this paper is to propose an IDS that use PCA and SVD algorithms to reduce the 42 IDS features and implement the reduced set to be recognized later by SVM classification algorithm.

By trial and error, the best k value is =21 which gives the highest accuracy when using PCA and SVD with SVM algorithm. It is clear from table (1), (2) and figure (3) that SVD algorithm gives better accuracy values than PCA algorithm. It is obvious from table(1),(2) and (6) that TP and TN values is much higher than FP and FN values which means that the proposed system gives good detection rates. finally the features for all traffic classes were successfully reduced with a feature selection algorithms PCA and SVD. This reduction is very important in minimizing the memory and CPU time.

REFERENCES

[1] Dasarathy, B.V.: Intrusion Detection, *Information Fusion*. (4) 243-245, 2003.

[2] American Computer Emergency Response Team /Coordination Center (CERT), http://www.cert.org/.

[3] Information Security Report, http://www.isecu-tech.com.tw/.

[4] Bace, R.G.: Intrusion Detection. Macmillan Technical Publishing. 2000.

[5] H. Debar etal. Towards a taxonomy of intrusion detection systems" Computer Network, pp. 805-822, April1999.

[6] KDDCup99dataset, August 2003 http://kdd.ics.uci.edu/databases/kddcup99 /kddcup9 9.html.

[7] R.O.Duda, P.E.Hart, and D.G.Stork, Pattern Classification, vol. 1. New York: Wiley, 2002.

[8] A.K.Jain, R.P.W.Duin, and J.Mao, "Statistical Pattern Recognition: A Survey," IEEE Transactions on Pattern Analysis and Mission Intelligence, vol. 22, pp.4-37, January 2000. Computer Science

[9] Abdoul Karim Ganame RB, Bourgeoisa Julien, Spiesa F. A global security architecture for intrusion detection on computer networks. Computers & Security pp. 30–47, 2008.

Feature Extraction Based Classification Technique For Intrusion Detection System 37

[10] Balasubramaniyan J, Garcia-Fernandez J, Isacoff D, Spafford E, Zamboni D. An architecture for intrusion detection using autonomous agents. In: Proceedings of the 14th IEEE computer security applications conference pp. 13–24, 1998.

[11] S. Peddabachigari, A. Abraham, C. Grosan and J. Thomas, Modeling Intrusion Detection Systems Using Hybrid Intelligent Systems. Journal of Network and Computer Applications, 30, pp. 114-132, 2007.

[12] A.N. Toosi, M. Kahani, A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers, Computer Communications pp. 2201–2212, 2007.

[13] Chung-Ming Ou, Host-based intrusion detection systems adapted from agent-based artificial immune systems, Neurocomputing pp. 78–86, 2012.

[14] H.G. Kayacik, A.N. Zincir-Heywood, M.I. Heywood, on the capability of an SOM based intrusion detection system, in: Proceedings of the 2003 International Joint Conference on Neural Networks, vol. 3, IEEE Press, pp. 432-444, 2003.

[15] Chenfeng Vincent Zhou, Christopher Leckie, Shanika Karunasekera, A survey of coordinated attacks and collaborative intrusion detection, computers & security 29, pp. 124–140, 2010

[16] D. Mutz, F. Valeur, G. Vigna and C. Kruegel. Anomalous system call detection. ACM Trans. Inf. Syst. Secur., volume 9, ISSN 1094-9224, pp. 61-93, 2006.

[17] Hochberg J, Jackson K, Stallings C, McClary JF, DuBois D, Ford J. Nadir: an automated system for detecting network intrusion and misuse. In: Proceedings of the 15th national computer security conference, pp. 235–48, 1993.

[18] Fenet S, Hassas S. A distributed intrusion detection and response system based on mobile autonomous agents using social insects communication paradigm. In: Proceedings of the First International Workshop on Security of Mobile Multiagent Systems (SEMAS), pp. 41-58, 2001.

[19] Chittineni Aruna and R. Siva Ram Pra. Experimental Evaluation and Result Discussion of Metamorphic Testing Automation Framework with Novel Algorithms. International Journal of Computer Engineering and Technology, 7(1), 2016, pp. 26-35.

[20] R.P.S. Manikandan and Dr. A.M. Kalpana, Design of Transactional Prediction using Plan Mine and Genetic Algorithms. International Journal of Computer Engineering and Technology, 7(6), 2016, pp. 50–54.

[21] Hung-Jen Liao, Chun-HungRichardLin, Ying-ChihLin, Kuang-YuanTung, Intrusion detection system: A comprehensive review, Journal of Network and Computer Applications 36, pp. 16–24, 2011